WEB MINING: RESEARCH AND PRACTICE

Web mining techniques seek to extract knowledge from Web data. This article provides an overview of past and current work in the three main areas of Web mining research—content, structure, and usage—as well as emerging work in Semantic Web mining.

As a large and dynamic information source that is structurally complex and ever growing, the World Wide Web is fertile ground for data-mining principles, or *Web mining*. The Web mining field encompasses a wide array of issues, primarily aimed at deriving actionable knowledge from the Web, and includes researchers from information retrieval, database technologies, and artificial intelligence. Since Oren Etzioni,[1] among others, formally introduced the term, authors have used "Web mining" to mean slightly different things. For example, Jaideep Srivastava and colleagues[2] define it as

> The application of data-mining techniques to extract knowledge from Web data, in which at least one of structure or usage (Web log) data is used in the mining process (with or without other types of Web data).

Researchers have identified three broad categories of Web mining:[2,3]

- *Web content mining* is the application of data-mining techniques to content published on the Internet, usually as HTML (semistructured), plaintext (unstructured), or XML (structured) documents.
- *Web structure mining* operates on the Web's hyperlink structure. This graph structure can provide information about a page's ranking[4] or authoritativeness[5] and enhance search results through filtering.
- *Web usage mining* analyzes results of user interactions with a Web server, including Web logs, clickstreams, and database transactions at a Web site or a group of related sites. Web usage mining introduces privacy concerns and is currently the topic of extensive debate.

We discuss some important research contributions in Web mining, with a goal of providing a broad overview rather than an in-depth analysis.

Web Content and Structure Mining

Some researchers combine content and structure mining to leverage the techniques' strengths. Although not all researchers agree to such a classification, we list research in these two areas together.

Fabrizio Sebastini[6] and Soumen Chakrabarti[7] discuss Web content mining techniques in detail, and Johannes Fürnkranz[8] surveys work in Web structure mining.

1521-9615/04/$20.00 © 2004 IEEE
Copublished by the IEEE CS and the AIP

PRANAM KOLARI AND ANUPAM JOSHI
University of Maryland, Baltimore County

Web as a Database

Early work in the area of Web databases focused on the Web's layered view, as suggested by Osmar Zaiane and colleagues.[9] Placing a layer of abstraction containing some semantic information on top of the semistructured Web lets users query the Web as they would a database. For instance, users can readily query a metadata layer describing a document's author or topic. Researchers can use content and hyperlink mining approaches in which XML represents the semantics to build such a multilayered Web. WebLog[10] and WebSQL[11] are such database-based approaches. More recent work in this area aims to realize the Semantic Web vision.[12]

Document Classification

Classification's roots are in machine learning, pattern recognition, and text analysis. The basic idea is to classify pages using supervised or unsupervised methods. In simple terms, supervised learning uses preclassified training data, which is not required in unsupervised learning. Classification is useful in such areas as topic aggregation and Web-community identification.

Early work in document classification applied text-mining techniques to Web data directly. (Text mining is a subcategory of Web content mining that does not use Web structure.) Later research showed that harnessing the Web graph structure and semistructured content in the form of HTML tags improved results. Hypursuit is an early effort in this direction.[13] Google News (http://news.google.com), which automatically gathers and classifies the most recent news from more than 4,000 sources, is a popular application of document classification.

Hubs and Authorities

Hyperlink-induced topic search (HITS) is an iterative algorithm for mining the Web graph to identify topic *hubs* and *authorities*. "Authorities" are highly ranked pages for a given topic; "hubs" are pages with links to authorities. The algorithm takes as input search results returned by traditional text-indexing techniques, and filters these results to identify hubs and authorities. The number and weight of hubs pointing to a page determine the page's authority. The algorithm assigns weight to a hub based on the authoritativeness of the pages it points to. For example, a page containing links to all authoritative news servers (CNN, CNBC, and so on) is a powerful news hub.

Larry Page and colleagues proposed PageRank[4] and popularized it through the Google search engine. With PageRank, a crawler precomputes page ranks, increasing the speed with which ranked search results are returned. A page's PageRank computation is based on the number of links other ranked pages have to it and the probability that a surfer will visit it without traversing links (through bookmarks, for example). Researchers have suggested enhancements to the basic PageRank algorithm. Sepandar Kamwar and colleagues,[14] for example, developed a quadratic extrapolation algorithm that significantly improves the cost of PageRank computation.

Clever: Ranking by Content

Basic hub and authority approaches do not consider a link's semantics for page ranking. The Clever[15] system addresses this problem by considering query terms occurring in or near the anchor text (a certain window) in an HTML page as a hint to link semantics, and thus leverages content-mining techniques for structure analysis. Clever gives greater weight to links that are similar to the search query. It incorporates a link's weight into the HITS algorithm when deciding a page's authoritativeness. For example, if n pages link to two other pages for different reasons, such as business and sports, the enhanced HITS algorithm will return different ranks for both pages for queries on sports and business.

Soumen Chakrabarti and colleagues suggested additional refinements,[15] and their results show a significant improvement over contemporary approaches.

Identifying Web Communities

Many communities are well organized on the Web, with *webrings* (interlinks between Web sites with a ring structure) or information portals linking them together. Ravi Kumar and colleagues[16] proposed *trawling* to identify nascent or emerging communities using hyperlink data to obtain cocitation information. They represent such a group or community as a dense directed bipartite graph with nodes divided into the community core and the rest. The community core represents those Web sites that are part of the same community without links between themselves. Trawling is the process of identifying such subgraphs from the Web graph.

Web Usage Mining

Web usage mining has several applications in e-business, including personalization, traffic analysis, and targeted advertising. The development of graphical analysis tools such as Webviz[17] popularized Web usage mining of Web transactions. The main areas of research in this domain are Web log data preprocessing and identification of useful patterns from this preprocessed data using mining techniques. Several

surveys on Web usage mining exist.[18,19]

Most data used for mining is collected from Web servers, clients, proxy servers, or server databases, all of which generate noisy data. Because Web mining is sensitive to noise, data cleaning methods are necessary. Jaideep Srivastava and colleagues[18] categorize data preprocessing into subtasks and note that the final outcome of preprocessing should be data that allows identification of a particular user's browsing pattern in the form of page views, sessions, and clickstreams. Clickstreams are of particular interest because they allow reconstruction of user navigational patterns. Recent work by Yannis Manolopoulos and colleagues[20] provides a comprehensive discussion of Web logs for usage mining and suggests novel ideas for Web log indexing. Such preprocessed data enables various mining techniques. We briefly describe some of the notable research here.

Adaptive Web Sites

Personalization is one of the most widely researched areas in Web usage mining. An early effort in this direction was the adaptive Web site challenge posed by Oren Etzioni and colleagues.[21] Adaptive sites automatically change their organization and presentation according to the preferences of the user accessing them. Other contemporary research seeks to build agent-based systems that give user recommendations. For instance, Web-watcher[22] uses content- and structure-mining techniques to give guided tours to users browsing a page. Popular Web sites like Amazon.com use similar techniques for "recommended links" provided to users. All these approaches primarily use association rules and clustering mechanisms on log data and Web pages.

Robust Fuzzy Clustering

Anupam Joshi and colleagues[23] use fuzzy techniques for Web page clustering and usage mining, and they use the mined knowledge to create adaptive Web sites.[24] They argue that given the inherent ambiguity and complexity of the underlying data, clustering results should not be clearly demarcated sets but rather fuzzy sets—that is, overlapping clusters. For instance, a user can belong to multiple user interest groups because at different times he or she accesses the Web for different information or merchandise. Insisting that each user fit only a single group is clearly inconsistent with this reality.

Moreover, given the noise expected in the data despite cleaning attempts, the clustering process must be robust in the statistical sense. Raghu Krishnapuram and colleagues discuss fuzzy clustering and its application to Web-log analysis and present a fast linear clustering algorithm that can handle significant data noise.[24] They use this algorithm to cluster Web access logs and use the traversal patterns identified for specific groups to automatically adapt the Web site to those groups.

Association Rules

Early systems used collaborative filtering for user recommendation and personalization. Bamshad Mobasher and colleagues[25] used association-rule mining based on frequent item sets and introduced a data structure to store the item sets. They split Web logs into user sessions and then mined these sessions using their suggested association-rule algorithm. They argue that other techniques based on association rules for usage data do not satisfy the real-time constraints of recommender systems because they consider all association rules prior to making a recommendation. Ming-Syan Chen and colleagues[26] proposed a somewhat similar approach that uses a different frequent itemset counting algorithm.

Adaptive sites automatically change their organization and presentation according to the preferences of the user accessing them.

Recommender Systems

J. Ben Schafer and colleagues[27] note that recommender systems have enhanced e-business by

- converting browsers to buyers,
- increasing cross-sell by identifying related products, and
- building loyalty.

These systems primarily use association rule mining for pattern detection. In an e-business scenario, a recommender system uses customers' Web baskets (shopping carts) as data sources. Amazon.com has the most prominent application: "Customers who bought product A also bought product B."

Web Site Evaluation

Myra Spiliopoulou[28] suggests applying Web usage mining to Web site evaluation to determine needed modifications—primarily to the site's design of page content and link structure between pages. Such evaluation is one of the earliest steps

in Web usage analysis conducted by Web sites and is necessary for repeat visitors. Evaluation is important because all subsequent Web usage mining techniques are effective only in the presence of large amounts of data created by repeat visitors. The main technique for evaluating data is to model user navigation patterns and compare them to site designers' expected patterns. The Web utilization miner (WUM)[28] analysis tool, for example, incorporates evaluation.

Hamlet: To Buy or Not to Buy

Etzioni and colleagues[29] applied Web mining to airline ticket purchasing. Airlines use sophisticated techniques to manage yield, varying ticket prices according to time and capacity. Etzioni's approach mined airline prices available on the Web and price changes over time to produce recommendations regarding the best time to buy tickets. Many more innovative areas are yet to be explored.

Privacy Issues

Recent data-mining privacy violations have caused concern, specifically when data mining has involved vertically partitioned data—that is, data about the same entity or individual from multiple sources. One example is Terrorist Information Awareness (TIA, www.epic.org/privacy/profiling/tia), a DARPA-initiated program that aims to aggregate information from disparate sources to detect patterns that might indicate a terrorist. This program has led to serious public debate about whether such a system, even if technologically possible, should be used.

DoubleClick's (www.doubleclick.com) online advertising is an instance of tracking user behavior across multiple sites. If a user's transactions at every Web site are identified through uniquely identifiable information collected by Web logs, they could create a far more complete profile of the user's shopping habits. The current Web privacy architecture provided by the Platform for Privacy Preferences (P3P, www.w3.org/P3P) Protocol and A P3P Preference Exchange Language (APPEL) lets users control this kind of usage by explicitly agreeing or disagreeing to such tracking. However, Web sites and popular Web browsers offer limited support for such tools. Web mining research should accommodate this preference set and enforce it across organizations and databases. Lorrie Cranor surveys possible research direction in this area.[30]

Semantic Web Mining

The Semantic Web[12] is emerging as the next-generation Web, with a semantically rich language such as the Web Ontology Language (www.w3.org/TR/owl-features) for marking up hypertext pages. OWL allows more complex assertions about a page (for instance, its provenance, access rules, and links to other pages) than the Web-as-database approach, which is limited to simple metadata (topics, author, creation date, and so on). Moreover, these assertions will be in a language with explicit semantics, making it machine interpretable. As Bettina Berendt and colleagues[31] discuss, the Semantic Web and Web mining can fit together: Web mining enables the Semantic Web vision, and the Semantic Web infrastructure improves Web mining's effectiveness.

In the Semantic Web, adding semantics to a Web resource is accomplished through explicit annotation (based on an ontology). Humans cannot be expected to annotate Web resources; it is simply not scalable. Hence, we need to automate the annotation process through ontology learning, mapping, merging, and instance learning. Web content-mining techniques can accomplish this. For instance, we can use topic classification to automatically annotate Web pages with information about topics in an ontology. Annotations of this kind enable new possibilities for Web mining. Ontologies can help improve clustering results through feature selection and aggregation (for instance, identifying that two different URLs both point to the same airfare search engine). With the Semantic Web, page ranking is decided not just by the approximated semantics of the link structure, but also by explicitly defined link semantics expressed in OWL. Thus, page ranking will vary depending on the content domain. Data modeling of a complete Web site with an explicit ontology can enhance usage-mining analysis through enhanced queries and more meaningful visualizations.

Recent research has mostly focused on Web usage analysis, partly because of its applicability in e-business. We expect privacy issues, distributed Web mining, and Semantic Web mining to attract equal, if not more, interest from the research community. Increased use of Web mining techniques will require that privacy issues be addressed, however. Similarly, aggregating data in a central site and then mining it is rarely scalable, hence the need for distributed mining techniques. Finally, researchers will need to leverage the semantic information the Semantic Web provides. Exposing content semantics and the link explicitly can help in many tasks, including mining the hidden Web—that is, data stored in databases and not accessible through search engines.

As new data is published every day, the Web's utility as an information source will continue to grow. The only question is: Can Web mining catch up to the WWW's growth? CiSE

Acknowledgments

The US National Science Foundation helped support this work under grant NSF IIS-9875433, as did the DARPA DAML program under contract F30602-97-1-0215.

References

1. O. Etzioni, "The World Wide Web: Quagmire or Gold Mine?," *Comm. ACM*, vol. 39, no.11, 1996, pp. 65–68.
2. J. Srivastava, P. Desikan, and V. Kumar, "Web Mining: Accomplishments and Future Directions," *Proc. US Nat'l Science Foundation Workshop on Next-Generation Data Mining* (NGDM), Nat'l Science Foundation, 2002.
3. R. Kosala and H. Blockeel, "Web Mining Research: A Survey," *ACM SIGKDD Explorations*, vol. 2, no. 1, 2000, pp. 1–15.
4. L. Page et al., *The PageRank Citation Ranking: Bring Order to the Web*, tech. report, Stanford Digital Library Technologies, 1999-0120, Jan. 1998.
5. J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Proc. 9th Ann. ACM–SIAM Symp. Discrete Algorithms*, ACM Press, 1998, pp. 668–677.
6. F. Sebastini, *Machine Learning in Automated Text Categorization*, tech. report B4-31, Istituto di Elaborazione dell'Informatione, Consiglio Nazionale delle Ricerche, Pisa, 1999.
7. S. Chakrabarti, "Data Mining for Hypertext: A Tutorial Survey," *ACM SIGKDD Explorations*, vol. 1, no. 2, pp. 1–11, 2000.
8. J. Fürnkranz, "Web Structure Mining: Exploiting the Graph Structure of the World Wide Web," *Österreichische Gesellschaft für Artificial Intelligence* (ÖGAI), vol. 21, no. 2, 2002, pp. 17–26.
9. O.R. Zaïane and J. Han, "Resource and Knowledge Discovery in Global Information Systems: A Preliminary Design and Experiment," *Proc. 1st Int'l Conf. Knowledge Discovery and Data Mining* (KDD), AAAI Press, 1995, pp. 331–336.
10. L.V.S. Lakshmanan, F. Sadri, and I.N. Subramanian, "A Declarative Language for Querying and Restructuring the Web," Proc. *6th IEEE Int'l Workshop Research Issues in Data Eng., Interoperability of Nontraditional Database Systems* (RIDE-NDS), IEEE CS Press, 1996.
11. A. Mendelzon, G. Michaila, and T. Milo, "Querying the World Wide Web," *Proc. 1st Int'l Conf. Parallel and Distributed Information System*, IEEE CS Press, 1996, pp. 80–91.
12. T. Bemers-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific Am.*, vol. 279, no. 5, 2001, pp. 34–43.
13. R. Weiss et al., "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering," *Proc. 7th ACM Conf. Hypertext*, ACM Press, 1996.
14. S.D. Kamvar et al., "Extrapolation Methods for Accelerating PageRank Computations," *Proc. 12th Int'l World Wide Web Conf.*, ACM Press, 2003.
15. S. Chakrabarti et al., "Mining the Web's Link Structure," *Computer*, vol. 32, no. 8, 1999, pp. 60–67.
16. R. Kumar et al., "Trawling the Web for Emerging Cybercommunities," *Proc. 8th World Wide Web Conf.*, Elsevier Science, 1999.
17. J.E. Pitkow and K. Bharat, "WebViz: A Tool for WWW Access Log Analysis," *Proc. 1st Int'l Conf. World Wide Web*, Elsevier Science, 1994, pp. 271–277.
18. J. Srivastava et al., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *ACM SIGKDD Explorations*, vol. 1, no. 2, 2000, pp. 12–23.
19. M Eirinaki and M Vazirgiannis, "Web Mining for Web Personalization," *ACM Trans. Internet Technology*, vol. 3, no. 1, 2003, pp. 1–27.
20. Y. Manolopoulos et al., "Indexing Techniques for Web Access Logs," *Web Information Systems*, IDEA Group, 2004.
21. M. Perkowitz and O. Etzioni, "Adaptive Web Site: An AI Challenge," *Proc. Int'l Joint Conf. Artificial Intelligence* (IJCAI), Morgan Kaufmann, 1997.
22. R. Armstrong et al., "Webwatcher: A Learning Apprentice for the World Wide Web," *Proc. AAAI Spring Symp. Information Gathering from Heterogeneous, Distributed Environments*, AAAI Press, 1995, pp. 6–13.
23. A. Joshi and R. Krishnapuram, "Robust Fuzzy Clustering Methods to Support Web Mining," *Proc. ACM SIGMOD Workshop Data Management and Knowledge Discovery*, ACM Press, 1998.
24. T. Kamdar, *Creating Adaptive Web Servers Using Incremental Weblog Mining*, masters thesis, Computer Science Dept., Univ. of Maryland, Baltimore, C0–1, 2001.
25. B. Mobasher et al., "Effective Personalization Based on Association Rule Discovery from Web Usage Data," *Proc. 3rd ACM Workshop Web Information and Data Management* (WIDM 2001), ACM Press, 2001, pp. 9–15.
26. M.-S. Chen, J.S. Park, and P.S. Yu., "Efficient Data Mining for Path Traversal Patterns," *IEEE Trans. Knowledge and Data Eng.*, vol. 10, no. 2, 1998, pp. 209–221.
27. J.B. Schafer, J. Konstan, and J. Riedl, "Electronic Commerce Recommender Applications," *J. Data Mining and Knowledge Discovery*, vol. 5, nos. 1/2, 2000, pp. 115–152.
28. M. Spiliopoulou, "Web Usage Mining for Site Evaluation," *Comm. ACM*, , vol. 43, no. 8, 2000, pp. 127–134.
29. O. Etzioni et al., "To Buy or Not to Buy: Mining Airline Fare Data to Minimize Ticket Purchase Price," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, ACM Press, 2003.
30. L. Cranor, "I Didn't Buy It for Myself: Privacy and E-Commerce Personalization," *Proc. ACM Workshop Privacy in the Electronic Society*, ACM Press, 2003.
31. B. Berendt, A. Hotho, and G. Stumme, "Towards Semantic Web Mining," *Proc. US Nat'l Science Foundation Workshop Next-Generation Data Mining* (NGDM), Nat'l Science Foundation, 2002.

Pranam Kolari is a graduate student in the Computer Science Department at the University of Maryland, Baltimore County. His research interests include the Semantic Web, social network analysis, machine learning, and Web mining. Kolari has a BE in computer science from Bangalore University, India. He is a member of the ACM. Contact him at kolari1@cs.umbc.edu.

Anupam Joshi is an associate professor of computer science and electrical engineering at UMBC. His research interests are in the broad area of networked computing and intelligent systems, with a primary focus on data management for mobile computing systems, and most recently on data management and security in pervasive computing and sensor environments. He is also interested in the Semantic Web and data and Web mining. Joshi has a BTech in electrical engineering from the India Institute of Technology, Delhi, and an MS and PhD in computer science from Purdue University. He is a senior member of the IEEE and a member of the IEEE Computer Society and the ACM. Contact him at joshi@cs.umbc.edu.